Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Sašo Jakljevič

Sklop video predstavitev problemov v podatkovnem rudarjenju

diplomsko delo

visokošolski strokovni študijski program prve stopnje računalništvo in informatika

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Na področju podatkovnih znanosti so čedalje bolj priljubljena orodja, ki uporabljajo vizualno programiranje in uporabnikom omogočajo gradnjo tudi zahtevnejših postopkov analize podatkov brez uporabe programskih jezikov. Za izbrana odprtokodna orodja te vrste preglejte, kakšno podporo nudijo uporabnikom-začetnikom. Osredotočite se na video vsebine. Nato kot primer teh izdelajte izobraževalne motivacijske videe, ki predstavijo izbrane aspekte orodja Orange.

Kazalo

Povzetek

Naslov: Sklop video predstavitev problemov v podatkovnem rudarjenju

S porastom količine podatkov je področje podatkovnega rudarjenja vse bolj priljubljeno. Da bi uporabo orodij strojnega učenja, vizualizacij podatkov in iskanja zakonitosti v podatkih omogočili čim večji skupini uporabnikov, se v zadnjem času pojavljajo programi, ki uporabljajo tehnike vizualnega programiranja. S temi programi je moč postopke podatkovne analitike snovati vizualno, brez uporabe posebnih programskih jezikov. Primeri takih orodij so RapidMiner, KNIME, Weka in Orange. Za ta orodja v nalogi pregledamo, na kakšen način preko kratkih videov predstavljajo svojo funkcionalnost. Končni izdelek naše naloge sta nato dva motivacijska videa, ki predstavita primera podatkovnega rudarjenja in uporabo orodja Orange ter z njim nekaj izbranih tehnik vizualizacije podatkov in napovednih modelov.

Ključne besede: podatkovno rudarjenje, video predstavitev, izobraževalne vsebine, programska orodja .

Abstract

Title: Batch of video presentation of challanges in data mining

As we store more and more data, Data mining field grows increasingly popular. To enable usage of machine learning tools, visualizations and knowledge discovery in data to the widest possible audience, computer applications that leverage techniques of visual programing stated to appear. These applications enable visual structuring of data analytic processes without the usage of special programming languages. Examples of such tools are RapidMiner, KNIME, Weka and Orange. In the scope of this paper we look at how functionalities of these tools are presented. End result of the paper are two motivational videos, that present examples of data mining and usage of data mining application called Orange. Examples will show a few chosen techniques of visualization and prediction models.

Keywords: data mining, video presentation, educational content, software.

prazna stran

Poglavje 1

Uvod

Podatkovno rudarjenje je proces pridobivanja znanja, včasih imenovan tudi izkopavanje znanja (angl. knowledge discovery from databases) iz velikih količin podatkov, v njegovem osrčju pa so metode za odkrivanje vzorcev v podatkih [3]. Je interdisciplinarno področje, saj se med izvajanjem samega procesa pridobivanja znanja običajno srečujemo s podatkovnimi bazami in podatkovnimi skladišči, ki hranijo podatke, statističnimi metodami ocenjevanja in analize podatkov, algoritmi iz področja strojnega učenja, procesiranja signalov in slik, podatkovnih vizualizacij in še mnogih drugih področij [3, 2]. Poleg tega je potrebno razumeti tudi domeno iz katere izhajajo podatki, ki jih poskušamo pretvoriti v uporabno znanje.

Začetki podatkovnega rudarjenja segajo nazaj v 60. leta preteklega stoletja, ko se je področje podatkovnih baz začelo sistematično razvijati in se v 80. letih uveljavilo na trgu [3]. Prihod spleta in cenovno dosegljivih osebnih računalnikov ter drugih naprav je tekom let povečal ustvarjanje, pretok in analizo podatkov. Kopičeni podatki so začeli presegati človeške zmožnosti analize, ta podatkovno bogata, a informativno revna situacija, pa je poganjala razvoj področja podatkovnega rudarjenja in ga poganja še vedno [3].

Podatkovno rudarjenje je danes razširjeno na mnoga področja: vse od zdravstva, kjer se uporablja za razvoj novih farmacevtskih sredstev in terapij, raziskovanje človeškega genoma; na področju financ se uporablja za

zaznavanje zlorab kreditnih kartic in zavarovanj; trgovska industrija lahko do določene mere uspešno predvidi obnašanja kupcev in izboljša splošno nakupovalno izkušnjo, optimizira dobavo in transport blaga ter količine zalog. Podatkovno rudarjenje se je razširilo tudi na izobraževanje, in sicer za ugotavljanje, kako izboljšati načine poučevanja in pravočasno zaznavanje učencev, ki bi potrebovali pomoč. To so le nekatera področja in aplikacije med mnogimi, ki jih nisem omenil [7, 1].

Naslednji velik razmah v količini podatkov je predviden s prihodom interneta stvari (angl. Internet of things ali IoT), z njim pa tudi novi izzivi v podatkovnemu rudarjenju [3].

1.1 Orodja

S porastom količine podatkov in novimi izzivi se razvijajo in tem izzivom prilagajajo tudi orodja [8], ki so se na trgu pojavila tako v odprtokodni kot tudi v komercialni obliki. Med njimi so nekatera v bistvu programski jeziki in knjižnice za programske jezike, ki uporabnikom omogočajo veliko fleksibilnost pri podatkovnem rudarjenju, medtem ko druga v namene podpore vizualizacije in interakcije z uporabnikom uporabljajo grafične vmesnike, ki omejujejo fleksibilnost, a naredijo postopke pridobivanja znanja iz podatkov bolj enostavne [6, 5]. Vsaka vrsta orodij ima svoje prednosti in slabosti oz. pomanjkljivosti (orodje morda ne omogoča gradnje določenih vrst modelov, vizualizacij ali katere druge funkcije), zato izbira najprimernejšega orodja ni enostavna [6].

Glede na članek o priljubljenosti programskih orodij na `r4stats.com` so ključna vprašanja pri izbiri najprimernejšega orodja:

1. “Ali lahko teče na vašem osebnem računalniku?”

2. “Ali programska oprema podpira vse metode, ki jih potrebujete? Če ne, kako razširljiva je?”

3. “Ali njena razširljivost uporablja lasten jezik ali zunanjega (npr. Python, R), ki je široko dostopen preko dodatkov?”

4. “Ali v popolnosti podpira način uporabe (programiranje, meniji in pogovorna okna ali diagrami poteka oz. vizualno programiranje), ki vam ustreza?”

5. “Ali so možnosti vizualizacije (npr. statične ali interaktivne) zadovoljive za vaše probleme?”

6. “Ali omogoča izvoz v obliki, ki vam ustreza (npr. kopiraj in prilepi v program za procesiranje besedil ali pa integracija za LaTeX)?”

7. “Ali lahko procesira dovolj velike nabore podatkov?”

8. “Ali to orodje uporabljajo tudi vaši sodelavci, da lahko na enostaven način z njimi delite podatke in programe?”

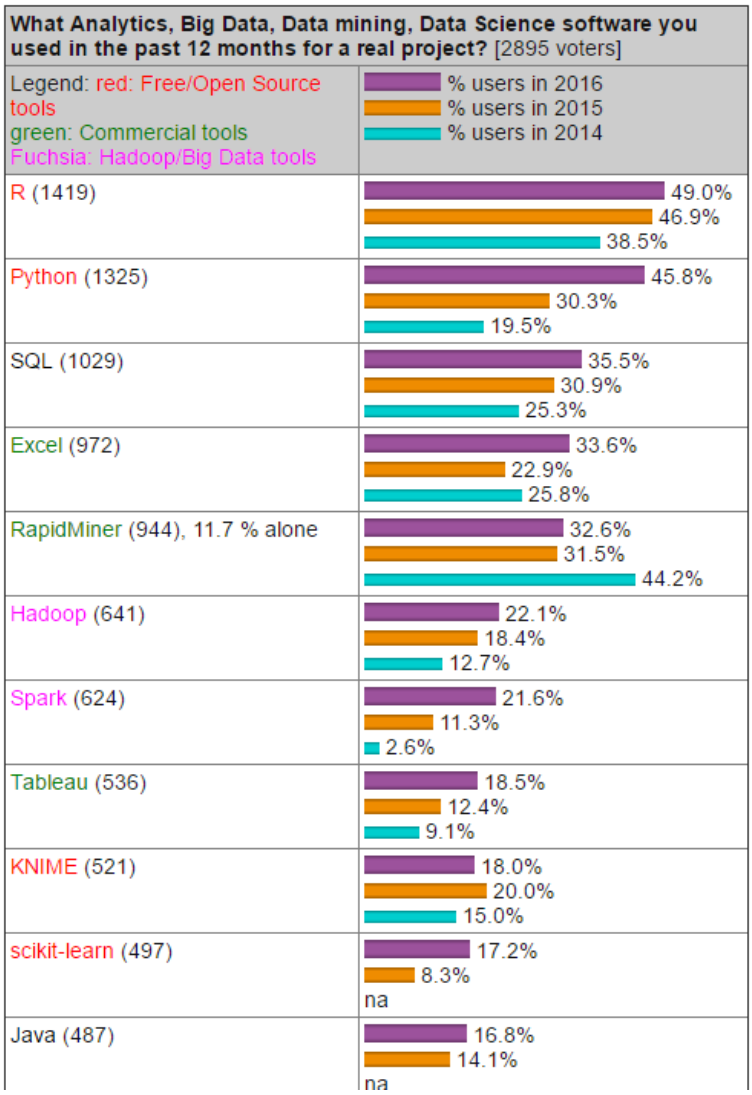
9. “Ali je v vašem finančnem dosegu [6]?”

Na spletni strani KDNuggets[1] so svoje obiskovalce, ki izhajajo iz krogov analitike in znanosti o podatkih, vprašali “Katero programsko opremo ste uporabili za analizo, podatkovno rudarjenje, podatkovno znanost, strojno učenje v projektih tekom zadnjih 12 mesecih?”. V anketi za leto 2016 je več kot 2895 glasovalcev lahko izbiralo med 102 orodji [9].

Iz slike 1.1, ki vsebuje najvišje uvrščena orodja v anketi, je razvidno, da na vrhu ostaja programski jezik R, sledi mu Python, za tem srečamo SQL in Excel, na petem mestu pa se znajde prvo orodje z grafičnim uporabniškim vmesnikom za vizualno programiranje, RapidMiner [9].

Problem tovrstnih anket je nekontroliran vzorec in posledično potencialno narobe utežen rezultat, ki lahko prikaže večjo ali manjšo uporabnost nekega orodja [6]. Nekoliko bolj celovito sliko nam lahko da pogled iz več perspektiv, česar se je lotil Robert A. Muenchen, ki se je pri svoji analizi omejil na 27

[1] `http://www.kdnuggets.com/`



Slika 1.1: Rezultati KDNuggets ankete za leto 2016, ki prikazuje odstotke glasov za najvišje uvrščena orodja. Za primerjavo so dodani tudi rezultati anket iz leta 2015 in 2014.

orodij in jezikov za analizo podatkov, za katere je zbral statistične podatke iz virov, kot so oglasi za službe, strokovni članki, ankete, knjige, blogi, forumi in še nekaj drugih.

Ugotovil je, da se med programsko opremo, ki se uporablja v obliki predpripravljenih metod za analizo, na vrhu vedno znajde R, SAS, SPSS in Stata. Med orodji, ki se uporabljajo kot programski jeziki, so na vrhu C/C#/C++, Java, MATLAB, Python, R in SAS [6]. Zaznal je tudi trend širitve orodij, ki uporabljajo grafični uporabniški vmesnik za vizualno programiranje, kot njihovo prednost pred orodji s klasičnimi grafičnimi uporabniškimi vmesniki pa je podal predvsem možnost shranjevanja in ponovne uporabe preteklih ustvarjenih projektov pri novih raziskovalnih nalogah. Njihova širitev omogoča uporabo močnih naprednih orodij za podatkovno analitiko uporabnikom, ki niso vešči programiranja [6]. Ravno zato se bomo v nadaljevanju naloge osredotočili na prav ta orodja, ki so hkrati tudi odprtokodna oz. brezplačna za uporabo in s tem prosto dostopna.

Orodja, čeprav imajo relativno preproste vmesnike, za uporabo niso enostavna, zato bomo analizirali, kako so ta orodja in načini uporabe predstavljeni na njihovih spletnih straneh in kakšne so njihove izobraževalne video vsebine. Glede na anketo spletne strani KDNuggets v letu 2016 so štiri najbolj znana tovrstna orodja, razvrščena glede na popularnost:

1. RapidMiner[2],
2. KNIME[3],
3. Weka[4] in
4. Orange[5] [9].

Kot izdelek naloge nas zanima, ali je mogoče, da začetniki v podatkovnem rudarjenju sestavijo video, ki bi širši populaciji približal problem podatkovne

[2] `https://rapidminer.com/`
[3] `https://www.knime.org/`
[4] `http://www.cs.waikato.ac.nz/ml/weka/`
[5] `http://orange.biolab.si/`

analitike in pomagal drugim na morda nekoliko neobičajen način približati to področje. Nalogo v nadaljevanju sestavljajo še 3 poglavja. V Poglavju 2 začnemo s pregledom orodij RapidMiner, KNIME, Weka in Orange ter analiziramo prisotnost njihovih izobraževalnih vsebin. V Poglavju 3 predstavimo pripravo predstavitvenih videov za orodje Orange, v Poglavju 4 pa povzamemo ugotovitve analize in možne izboljšave izdelka.

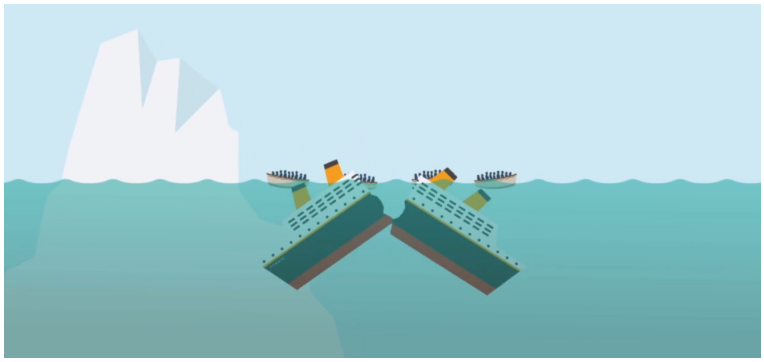
Poglavje 2

Analiza video izobraževalnih vsebin orodij za podatkovno rudarjenje

Podatkovno rudarjenje ima velik vpliv na mnogih področjih, ki višajo kakovost naših življenj in pripomorejo k tehnološkemu napredku [7]. S širitvijo znanj uporabe orodij za podatkovno rudarjenje na uporabnike, specializirane na raznih področjih in brez programerskih znanj, bi lahko ta vpliv še povečali ali pa vsaj seznanili potencialne bodoče uporabnike z uporabnostjo teh orodij [6]. V nadaljevanju analiziramo izobraževalne in predstavitvene vsebine s poudarkom na videu za štiri odprtokodna orodja z vmesnikom za vizualno programiranje, saj te vsebine predstavljajo enega od pomembnih načinov širjenja znanja [10].

2.1 RapidMiner

RapidMiner je odprtokodna platforma za analizo podatkov, ki zajema celoten proces pridobivanja znanja, omogoča pa tudi njegovo avtomatizacijo. Podpira veliko različnih algoritmov, omogoča razširitev za programska jezika R in Python in povezavo z orodji za procesiranje masovnih podatkov (angl. Big



Slika 2.1: Grafična animacija Titanika v predstavitvenem videu.

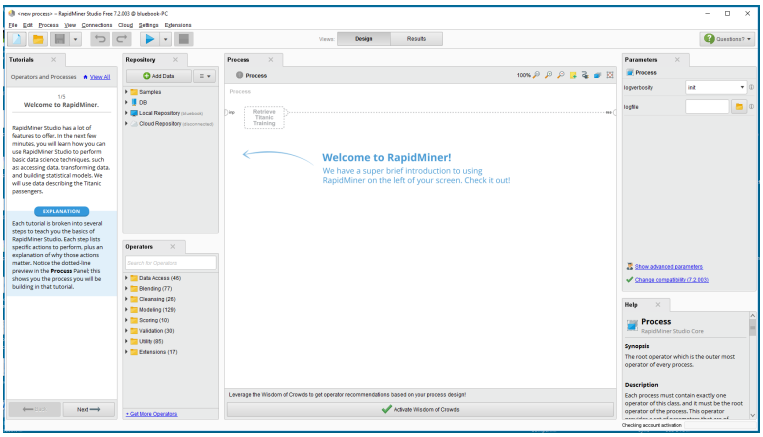
data), kot sta Hadoop[1] in Spark[2]. Znotraj programskega orodja se srečamo z vizualnim programiranjem. Ta omogoča vse od priprave podatkov, gradnje modelov do njihove implementacije.[3]

2.1.1 Prisotnost predstavitvenih in izobraževalnih video vsebin

S prvim videom, ki poskuša prepričati obiskovalca, kako hitro in enostavno je priti do prvih rezultatov ob uporabi RapidMiner-ja, se srečamo po kliku na prenos namestitvene datoteke. Tako si ga lahko ogledamo med prenosom programa na računalnik. Preko govorne razlage in enostavne grafične animacije, kot je vidna na sliki 2.1, je predstavljen učni primer Titanik. Postavi vprašanje, ali lahko v manj kot 30 sekundah dokažemo, da je bilo preživetje ob nesreči Titanika zgolj srečno naključje.

Za tem se premaknemo v program RapidMiner, kjer so predstavljeni zbrani podatki, govoru in sliki pa se pridružijo še grafični elementi, ki usmerjajo pozornost, ko je slika statična. V nekaj klikih je zgrajen model, video pa se zaključi s predstavitvijo ugotovitve, da so imele največjo verjetnost za preživetje ženske z otroki in dražjimi vstopnicami. S tem pokaže navidezno

[1] `http://hadoop.apache.org/`
[2] `http://spark.apache.org/`
[3] `https://rapidminer.com/`



Slika 2.2: Navodila za opravljanje uvodnih primerov znotraj aplikacije RapidMiner.

enostavnost in hitrost podatkovnega rudarjenja v orodju RapidMiner, ki je promovirana kot ena od njegovih največjih prednosti. Video pa ni le za namen promocije, ampak je tudi osnova, na podlagi katere se nadaljuje vodeni izobraževalni material.

Svoje začetno izobraževanje lahko nadaljujemo znotraj same aplikacije po korakih s pomočjo tekstovnih navodil in opornih grafičnih elementih ali pa na spletni strani ob pomoči kratkih, do 10-minutnih videov, v stilu drugega dela uvodnega videa. Vsebine nas peljejo skozi osnove RapidMinerja.

Znotraj aplikacije nam je po registraciji ponujeno vodeno delo. Kot vidimo na sliki 2.2, modra puščica usmeri pogled na levo stran okna, kjer se nahajajo navodila, prvi dve zapisani vaji pa nas zapeljeta skozi akcije v predstavitvenem videu. Znotraj programa je poleg teh dveh še 6 osnovnih vaj, nadaljujejo se s še 6 vajami na temo upravljanja s podatki (angl. data handling) in 5 na temo modeliranja, ocenjevanja in validacije (angl. modeling, scoring and validation). Najdemo tudi kratke neme izseke iz videov, ki predstavijo glavne akcije v orodju in neposredne povezave do ostalih vsebin na spletu.

Enaka vsebina, kot je predstavljena v obliki tekstovnih navodil za vajo, je

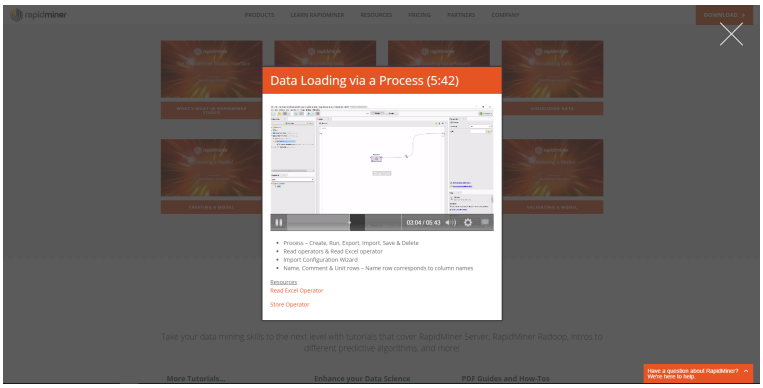
predstavljena tudi v obliki videov na spletni strani pod razdelkom "Getting started central" oz. "Learn RapidMiner". Videi predstavijo:

1. uporabniški vmesnik RapidMiner-ja;
2. vnos podatkov;
3. nalaganje podatkov preko procesa;
4. vizualizacijo podatkov;
5. ustvarjanje modela;
6. aplikacijo modela;
7. testiranje modela in
8. validacijo modela.

Ob kliku na povezavo do videa se odpre pojavno okno, kot ga vidimo na sliki 2.3. Pod videom vidimo povzetek glavnih segmentov v videu in seznam povezav do navezujoče se dokumentacije. Videi vsebujejo posnetek uporabniškega vmesnika, kjer ob kratki in jedrnati razlagi in dodanih grafičnih elementih prikazujejo, kako izvesti zadano nalogo.

Če smo že spoznali osnove, lahko na spletni strani najdemo še tri dodatne sklope video vsebin:

- Predstavitve naprednejših funkcionalnosti v RapidMinerju, kot je postavitev programa na strežniku ali pa delo z masovnimi podatki.
- "5 Minutes with Ingo", kakor se imenuje serija video predstavitev konceptov in algoritmov podatkovnega rudarjenja s predavateljem Ingo Mierswa. Ta s pomočjo fizičnih elementov, kot so risanje ali kocke, vizualizira svojo razlago. Namen teh videov je jasna razlaga in gledalčevo razumevanje konceptov, zato ne vsebujejo grafičnih elementov orodja RapidMiner. Razlaga običajno traja manj kot 10 minut, vsak video pa se začne s kratko igrano sceno in izsekom iz znanega filma.



Slika 2.3: Video na temo nalaganja podatkov v RapidMiner, ki se predvaja v pojavnem oknu spletne strani.

- Video demonstracije uporabe RapidMinerja pri kompleksnejših problemih v podatkovnem rudarjenju. Te najdemo pod zavihkom "Resources" na spletni strani. V videih je prikazano tako delo z RapidMinerjem, kot tudi hitra razlaga konceptov podatkovnega rudarjenja, ki se tičejo problema, katerega rešujejo.

Vsi videi so poleg spletni strani naloženi tudi na spletnem portalu youtube.com (kanal RapidMiner, Inc.) in po sklopih združeni v predvajalne liste. Med predvajalnimi listami najdemo tudi vsaj dve takšni, ki vsebujeta videe iz drugih kanalov ("RapidMiner Wisdom 2016 - New York City" in "e-LICO").

Poleg vsega naštetega ne smemo pozabiti na skupnost uporabnikov, mnoge članke, poročila, knjige in druge tovrstne vsebine ter možnost udeležbe plačljivih spletnih seminarjev, kjer se lahko izobražujemo na področju podatkovnega rudarjenja in uporabe RapidMinerja.

2.1.2 Povzetek

Orodje RapidMiner vsebuje zajetno količino predstavitvenih in poučnih video vsebin. Te so ustvarjene na način, da začetnika vodijo vse od prvih korakov uporabe RapidMinerja, nato skozi kompleksnejše probleme in

prikaz njihovega reševanja vse do konceptov, ki jih je treba upoštevati za uspešno podatkovno rudarjenje. Vsak sklop videov ima svoj segment, na katerega se osredotoča in je primeren za uporabnika s specifičnim predznanjem. Velika količina izobraževalnih gradiv naredi sam produkt zelo prijazen do začetnikov, ki niso programerji in se s podatkovnim rudarjenjem srečujejo prvič.

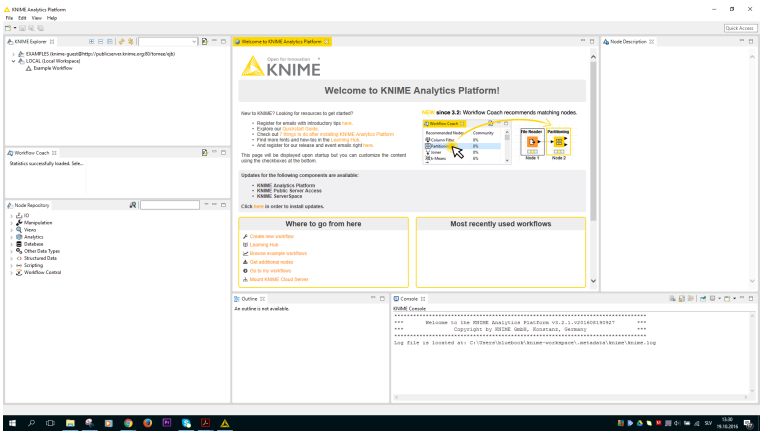
2.2 KNIME

KNIME je odprtokodna platforma za podatkovno rudarjenje napisana v Javi na osnovi Eclipse odprtokodnega razvojnega okolja. Velika skupnost prispeva k razvoju orodja, ki vsebuje več kot 1000 predpripravljenih modulov. Ti omogočajo vse od vnosa bolj priljubljenih podatkovnih formatov in povezav na podatkovne baze, njihovo združevanje znotraj istega procesa, do funkcij in naprednih algoritmov za strojno učenje, gradnjo napovednih modelov in analizo podatkov. Glavni način interakcije je grafični uporabniški vmesnik za vizualno programiranje, podpira pa tudi klasično programiranje modulov v več programskih jezikih (Python, R, Java, in mnogi drugi). Ta funkcionalnost za mnoga podjetja pomeni enostavnejši prenos že implementiranega znanja v KNIME in hkrati večjo fleksibilnost brez potrebe po učenju novega programskega jezika za njihove zaposlene.[4]

2.2.1 Prisotnost predstavitvenih in izobraževalnih video vsebin

KNIME nima izpostavljenega kratkega predstavitvenega videa, ki bi na enostaven način predstavil orodje. Nekoliko daljšo predstavitev lahko najdemo na YouTube kanalu, ki si ga bomo ogledali kasneje. Obisk spletne strani pokaže pisne vsebine, ki jih je na strani veliko. Ob kliku na povezavo za prenos pristanemo na strani s tremi koraki. Prvi korak je neobvezna registracija,

[4] `https://www.knime.org/`



Slika 2.4: Glavno okno delovnega okolja ob prvem zagonu programa KNIME z listo povezav. Te kažejo na stran za registracijo na mesečna e-poštna obvestila, vodič za hiter začetek uporabe orodja, članek s 7 nasveti, ki jih je dobro implementirati po inštalaciji orodja, in pa učno središče.

ki uporabnika prijavi na mesečno obveščanje z novicami in nasveti za nove uporabnike. Drugi korak je dejanski prenos programskega orodja, na tretjem pa je usmeritev uporabnika na podporne in izobraževalne vsebine:

- inštalacija in zagon programskega orodja KNIME;
- navodila, kako sestaviti prvi proces;
- povezava do učnega središča;
- povezava do zbirke primerov uporabe orodja;
- povezava na skupnostni forum.

Podoben seznam povezav uporabnik sreča tudi ob zagonu aplikacije KNIME (slika 2.4) v glavnem oz. srednjem oknu delovnega okolja.

Medtem ko je proces namestitve orodja predstavljen v obliki videa, je osnovna predstavitev uporabe orodja in prvi prikaz sestavljanja delovnih procesov v obliki besedila in fotografij.

Z večjo količino vsebin se srečamo v zbirki učnih gradiv (angl. Learning hub), katero vidimo na sliki 2.5. Ta je z zavihki ločena na:

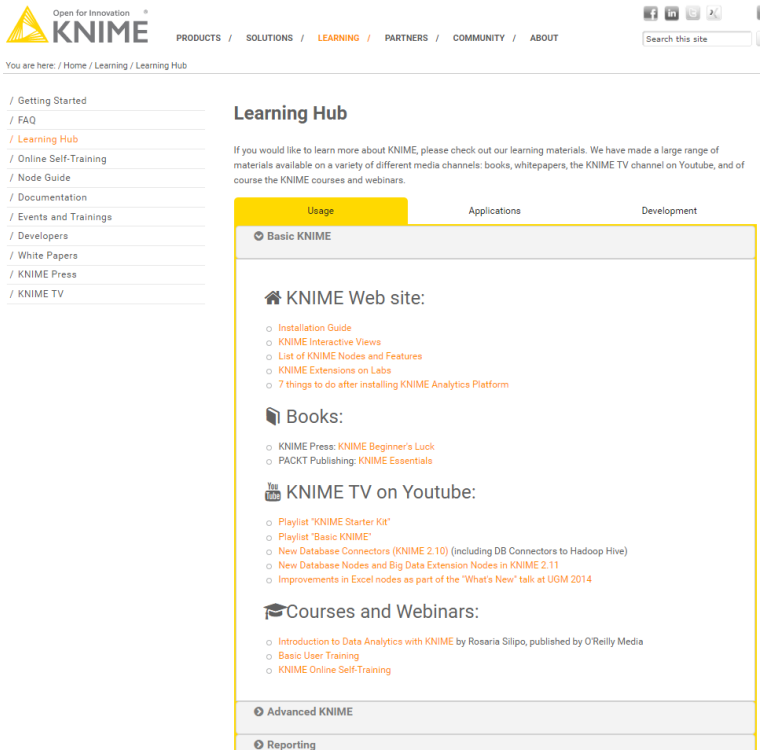
- uporabo orodja (angl. usage);
- aplikacijo orodja na probleme (angl. applications) in pa
- razvoj orodja (angl. development).

Zavihek za uporabo orodja vsebuje sekcije, razdeljene glede na nivo znanja in kompleksnosti uporabe orodja ("Basic KNIME", "Advanced KNIME", "Reporting", "KNIME Server"). Zavihek za aplikacijo orodja svojo vsebino razdeli po različnih domenah ("Data Mining", "Web, Text and Network Analysis", "Image Processing", "Chemistry in KNIME", "R, SAS, and Python", "Big Data", "Whitepapers"). Vsaka od podsekcij vsebujejo povezave na druge podstrani ali strani, knjige, videe in predvajalne liste, spletne seminarje, članke in poglobljena poročila. Zadnji zavihek o razvoju orodja vsebuje le eno podsekcijo s povezavo na napotke za razvijalce.

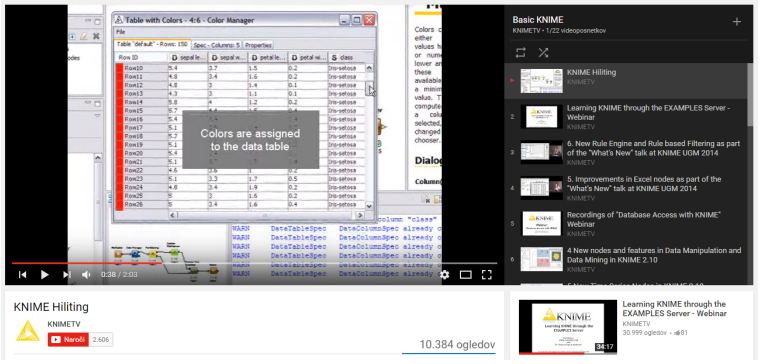
Zbirka primerov uporabe orodja KNIME (angl. Use Cases) vsebuje praktične primere s predlogami vizualnega procesa (angl. workflow) za razna področja. Podobno kot zbirka primerov na spletni strani, je tudi velikokrat omenjena in priporočena zbirka primerov na strežniku, do katerih lahko uporabnik dostopa neposredno iz aplikacije, če je registriran.

Med preostalimi učnimi vsebinami je vredno omeniti še "Online Self-training". Je brezplačni spletni seminar, razdeljen na lekcije, ki predstavijo, kako je videti proces podatkovnega rudarjenja od začetka do konca znotraj orodja KNIME (vključno z instalacijo orodja in postavitev strežnika). V lekciji najdemo tekstovne razlage, namensko posnete video učne vsebine kot tudi posnetke seminarjev in predstavitev novosti orodja ob izidu nove verzije. Obsegajo kombinacijo razlag konceptov in uporabe orodja, vsaka lekcija pa ima na koncu tudi povezave za nadaljnje branje, spisek vaj in vprašanja, s katerimi lahko uporabnik preveri svoje znanje.

Vsi videi se nahajajo na KNIME TV kanalu spletnega portala YouTube. Kanal gostuje posnetke spletnih seminarjev, predstavitev posameznih funkci-



Slika 2.5: Zbirka učnih gradiv (angl. Learning hub) na spletni strani programa KNIME. Na vrhu vidimo glavno navigacijo, na levi pa so povezave na druge učne vsebine.



Slika 2.6: Video predstavi označevanje podatkov v orodju KNIME. Na desni vidimo naslednje videe na predvajalni listi imenovani Basic KNIME. Naslednji video je posnetek spletnega seminarja.

onalnosti ob izhodu nove verzije orodja, predstavitve posameznih konceptov podatkovnega rudarjenja in klasične predstavitve reševanja problemov podatkovnega rudarjenja z orodjem KNIME.

Kakovost zvoka in slike je med videi zelo različna, kot tudi pristop predavatelja oz. govorca. Nekateri videi so nemi in vsebujejo le besedilno razlago v oblačkih, kar pomeni, da mora gledalec svojo pozornost prenašati med branjem in opazovanjem izvajanja akcij. Drugi videi vsebujejo besedilo in govorno razlago s popolnoma enako vsebino.

Videi dogodkov so pogosto razdeljeni na več delov, ti pa so združeni v predvajalne sezname kot na primer “Talks at KNIME UGM 2015”. Predvajalni seznami pa lahko vsebujejo tudi posnetke, ki obsegajo določeno temo ali nivo znanja kot na primer “KNIME Starter Kit”, “KNIME Basic” in “KNIME Advanced”. Slednji seznami so sestavljene iz videov različnih seminarjev in predstavitev, posledično pa njihova vsebina ni neposredno povezana. Na sliki 2.6 vidimo predstavitev označevanja podatkov v predvajalnem seznamu Basic KNIME.

2.2.2 Povzetek

KNIME ima veliko učnih vsebin, ki so na voljo njihovim uporabnikom. Med njimi je tudi veliko video vsebin z različnimi nivoji kakovosti slike in zvoka. Učne predvajalne liste vsebujejo videe, ki niso bili namensko narejeni za sistematično učenje, kjer je snov podana v nekem vrstnem redu, zato je tudi sledenje gledalca nekoliko oteženo.

2.3 Weka

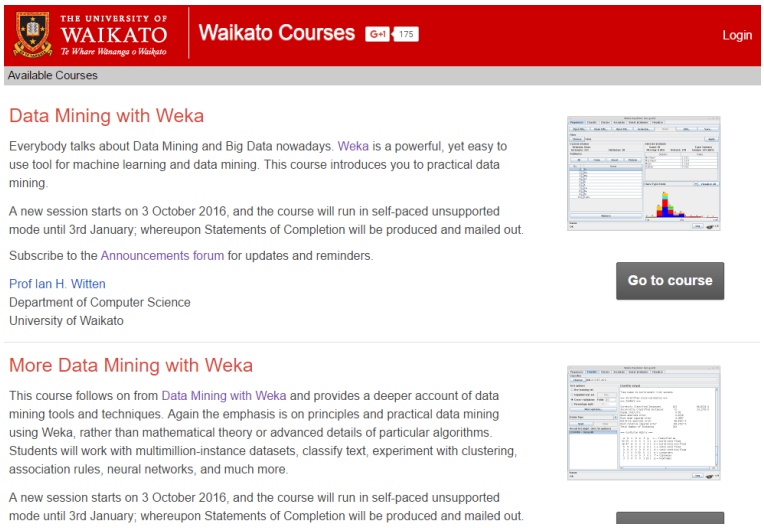
WEKA je odprtokodno delovno orodje (objavljeno pod GNU Licenco) za podatkovno rudarjenje, razvito s strani skupine za strojno učenje Wakiato Univerze na Novi Zelandiji. Vsebuje zbirko algoritmov, ki jo lahko uporabljamo preko štirih različnih uporabniških vmesnikov. Weka 3.8 daje tudi dostop do paketov za porazdeljeno podatkovno rudarjenje, kar pomeni podporo za delo z masovnimi podatki ter ogrodjema Hadoop in Spark.[5]

2.3.1 Prisotnost predstavitvenih in izobraževalnih video vsebin

Uradna spletna stran, namenjena programskemu orodju Weka, je v bistvu spletna stran skupine za strojno učenje na univerzi Waikato, stran kot taka pa na prvi pogled vsebuje le minimalno količino informacij o programskem orodju samem.

Ob prenosu programa ni posebnega usmerjanja na izobraževalne vsebine. Na podstrani "Software" najdemo dve povezavi na temo izobraževanja. Prva je preusmeritev na Waikato Courses vidna na sliki 2.7, kjer se lahko prijavimo na 3-mesečni spletni tečaj podatkovnega rudarjenja. Tega lahko opravljamo v svojem tempu, a nas obvezuje časovna omejitev tri mesece, ko so dostopne dodatne vsebine in potekajo tedenske objave nalog preko Googlove infrastrukture Google groups. Trenutno so ponujeni trije nivoji tečajev. Udeležba

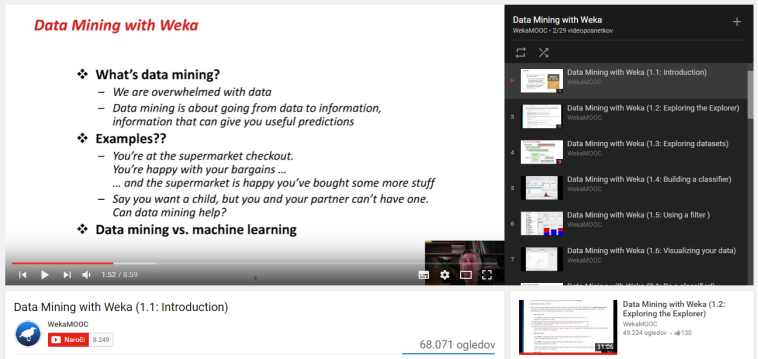
[5] `http://www.cs.waikato.ac.nz/ml/weka/index.html`



Slika 2.7: Spletna stran Waikato Courses, kjer se lahko prijavimo na 3-mesečni spletni tečaj podatkovnega rudarjenja.

tečaja na srednjem ali višjem nivoju nima predpogoja udeležbe tečaja na nižjem nivoju, je pa priporočljiva, saj se vsebina nadaljuje. Tečaji so brezplačni, njihov namen poleg učenja in promocije orodja WEKA pa je raziskava področja brezplačnih spletnih tečajev. Dvakrat med tečajem je vsak učenec deležen preverjanja znanja in če oba uspešno prestane, prejme za to potrdilo s strani Univerze Waikato.

Nekatere vsebine so dostopne tudi izven trajanja tečaja. Videe najdemo na WekaMOOC kanalu portala YouTube, kjer v predvajalnih listah najdemo združene posnetke za vse tri nivoje podatkovnega rudarjenja. Medtem ko vsebine za prva dva nivoja vodi le en predavatelj, se na tretjem nivoju pojavi več predavateljev iz Univerze Waikato, kjer vsak predstavlja svoje raziskovalno področje. Videi so sestavljeni iz kombinacije posnetka predavatelja, ki razlaga snov, ta pa je vizualno podprta s projekcijo, kot vidimo na sliki 2.8 in zajemom ekrana ob uporabi uporabniškega vmesnika Weka Exprorer. Videi so namensko ustvarjeni za določen vrstni red in imajo dosledno zadovoljivo kakovost slike in zvoka. Na koncu videa predavatelj povabi k opravljanju naloge, objavljene v času tečaja.



Slika 2.8: Prvi video v predvajalni listi najosnovnejšega tečaja Datamining with Weka.

Izobraževalne vsebine so tudi pod povezavo "Documentation", kjer je izpostavljen obširen priročnik uporabe programa, integriran v samo orodje, in mnoge povezave, ki so razdeljene na sklope:

- Splošna dokumentacija (angl. general documentation) - povezave na Wiki strani, liste elektronskih naslovov in njenega arhiva, dokumentacija skupnosti za orodje, povezave na priročnik, aplikacijski programski vmesnik (angl. application programming interface) in spisek dodatkov za programsko orodje.
- Razne informacije (angl. miscellaneous information) - z Weko povezan blog Marka Hallova, predstavitve in videi različnih načinov uporabe orodja in njegovih uporabniških vmesnikov, pa tudi predstavitev nastanka in usmeritve razvoja programa.
- Tehnične informacije (angl. technical information) - tehnične informacije, pomembne za napredne uporabnike za povezovanje z zunanjimi viri, uporabo različnih formatov in drugo.
- Priročniki starejših verzij programa Weka.

Na spletni strani je predstavljena tudi knjiga "Data Mining: Practical Machine Learning Tools and Techniques" v brezplačni prenos pa so ponu-

jene izobraževalne predstavitve na njeni osnovi, kot tudi povezave do zbirke člankov na temo podatkovnih znanosti. Za dodatna vprašanja ali dokumentacijo je na voljo tudi Wiki stran s forumom.

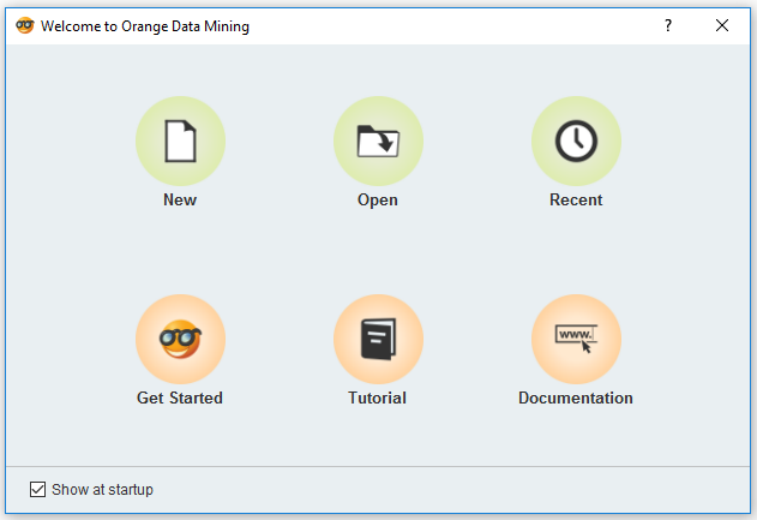
2.3.2 Povzetek

Kljub na videz zelo obskurni strani ta vsebuje premišljene izobraževalne vsebine, ki so hkrati brezplačne. Videi, pripravljeni za spletne tečaje uporabnika, strukturirano popeljejo od osnov uporabe orodja Weka do naprednih implementacij tehnik za pridobivanja znanja iz podatkov. V primeru prijave na tečaj in uspešno dokazanega znanja uporabnik prejme tudi dokazilo. Manjkajo vsebine, ki bi prikazale uporabo tudi ostalih uporabniških vmesnikov poleg Explorerja.

2.4 Orange

Orange je odprtokodno orodje za podatkovno analizo z grafičnim uporabniškim vmesnikom za vizualno programiranje. Ta začetnemu uporabniku olajša proces analize s tem, da usmeri njegovo pozornost v samo analizo in sestavljanje procesa namesto v programiranje rešitve, na drugi strani pa naprednejšim uporabnikom omogoča hitro pripravo prototipov procesov, ponuja pa tudi uporabo knjižnice za programiranje v programskem jeziku Python. Program vsebuje veliko orodij za analizo in vizualizacijo podatkov v obliki gradnikov (angl. widgets), z dodatki pa omogoča tudi povezovanje z zunanjimi podatkovnimi viri, procesiranje naravnega jezika, tekstovno rudarjenje in še kaj. V svojih procesih podpira celo gradnike, narejene specifično za učenje, kar ga naredi še bolj primernega za spoznavanje področja podatkovnega rudarjenja.[6]

[6] `http://orange.biolab.si/`



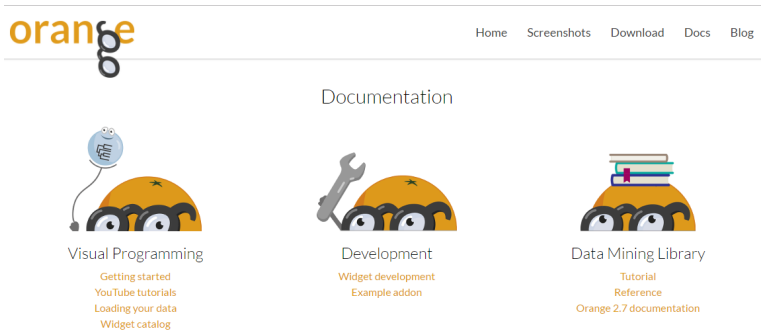
Slika 2.9: Začetno okno ob zagonu programa Orange. V spodnji vrstici vidimo povezave do izobraževalnih vsebin.

2.4.1 Prisotnost predstavitvenih in izobraževalnih video vsebin

Izobraževalne vsebine programa trenutno temeljijo predvsem na razlagah s pomočjo besedila in fotografij, s katerimi se lahko srečamo na več delih spletne strani. Za začetnega uporabnika je pripravljen začetni vodič ("Getting started"), ki ob prenosu datoteke ni posebej poudarjen. Nanj nas usmeri pojavno okno ob zagonu aplikacije Orange (vidno na sliki 2.9) ali pa se s povezavo do te strani srečamo ob obisku podstrani "Docs" oz. "Documentation".

Začetni vodič najprej predstavi pojavno okno (vidno na sliki 2.9) ob prvem zagonu in izpostavi možnost učenja preko primerov, shranjenih v sami aplikaciji skupaj z njihovimi razlagami (angl. tutorials). V nadaljevanju kombinacija besedila in slik prikazov uporabniškega vmesnika razloži osnove uporabe orodja Orange s primeroma Iris in Titanic.

Zbirka učnih primerov se nahaja tudi na podstrani "Screenshots" (povezavo najdemo v glavnem navigacijskem meniju). Za razliko od primerov v



Slika 2.10: Izobraževalne vsebine, zbrane na podstrani Docs za orodje Orange.

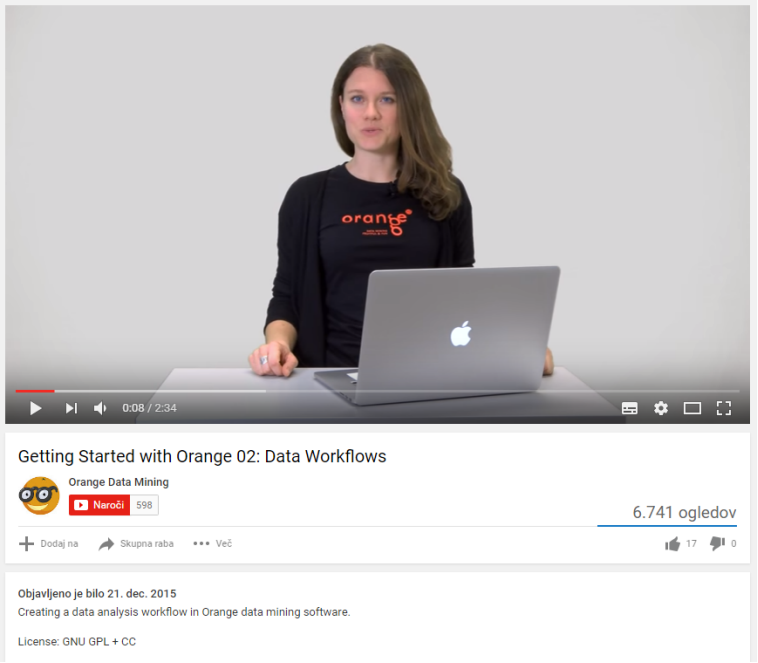
aplikaciji ti niso interaktivni, ampak so v obliki zajetih fotografij. Med njimi so predvsem različni načini vizualizacij, združevanja podatkov in grajenja modelov z različnimi podatki.

Če se vrnemo na izobraževalne vsebine pod "Docs" oz. "Documentation", vidne na sliki 2.10, so razdeljene na:

- vizualno programiranje - izobraževalne vsebine za uporabnike grafičnega vmesnika;
- razvoj - dokumentacija na temo razvoja gradnikov za Orange;
- knjižnica za podatkovno rudarjenje - vaje in dokumentacija za spoznavanje in lažje podatkovno rudarjenje v programskem jeziku Python s pomočjo knjižnice Orange.

Poleg prej omenjene "Getting started" strani, sekcija vizualnega programiranja vsebuje povezavo do YouTube kanala, povezavo do obširne dokumentacije o pripravi in nalaganju podatkov iz različnih virov in formatov, zadnja pa je katalog gradnikov z opisi delovanja in primeri uporabe.

YouTube kanal se imenuje Orange Data Mining in trenutno vsebuje 9 videov, ki trajajo do največ 5 minut. V njih predavateljica, ki jo vidimo na sliki 2.11, razlaga in preko zajema grafičnega vmesnika prikazuje upo-



Slika 2.11: Zajeta fotografija iz videa o ustvarjanju delovnih potekov (angl. workflows).

rabo orodja na primerih. Med razlago se prikaz dela v grafičnem vmesniku izmenjuje z videom predavateljice.

Skupek izobraževalnih vsebin se nahaja tudi med vsebinami Orange bloga, ki vsebuje mesečne objave na različne teme od primerov uporabe programa Orange in njegove knjižnice, do nasvetov za uporabo, dogodke, povezane z Orange-om, in njegove posodobitve. Blog omogoča pogled na zadnjih 5 objav, lahko iščemo po ključnih besedah ali pa si ogledamo skupek objav za vsak mesec posebej. Članki se pokažejo v celoti eden pod drugim glede na datum objave, izjema je le uporaba iskalnika ključnih besed. Takrat se članki prikažejo v skrčeni obliki, posledično pa lahko hitreje pregledamo naslove najdenih člankov.

2.4.2 Povzetek

Orange je med štirimi orodji, ki jih smo analizirali, orodje z najmanj video izobraževalnimi vsebinami, te, ki obstajajo, pa imajo zadovoljivo kakovost slike in zvoka. Vsebine si logično sledijo, a vsebinsko niso tako močno povezani, kakor pri RapidMinerju. Spletna stran ima tudi veliko tekstovnih in slikovnih razlag, a manjkajo vsebine, urejene v določen vrstni red, kateremu bi lahko popolni začetnik sledil in si postopoma nadgrajeval znanje.

2.5 Diskusija

Na kratko lahko povzamemo, da ima RapidMiner največ dobro razdelanih video vsebin, s pomočjo katerih se lahko njihovi uporabniki izobražujejo. KNIME ima videe nižje kakovosti kljub njihovi večji količini (razvidno iz tabele 2.1), hkrati pa so tudi slabše povezani med seboj, kar lahko pomeni težje učenje. Weka ima celosten pristop v obliki spletnega tečaja z videi, ki se med seboj dobro povezujejo, vendar so narejeni predvsem za uporabniški vmesnik Weka Explorer. Orange je šele na začetku svoje poti ustvarjanja video izobraževalnih vsebin, zato lahko izkoristi možnost opazovanja svojih tekmecev, ki so na to pot že zakorakali.

Glede na to, da so tovrstna orodja primerna za uporabnike, ki so začetniki na tem področju, jih je dobro tudi predstaviti na ta način. V tabeli 2.2 vidimo povzetek predstavitvenih oz. promocijskih videov orodij. Kot vidimo, je video RapidMinerja kratek in predstavi tako začetniku razumljiv problem kot samo orodje. KNIME se osredotoči predvsem na predstavitev funkcij, ki začetniku povedo bolj malo, hkrati pa je kljub zanimivi začetni animaciji kakovost nizka. Video orodja Weka je precej daljši kakor videi ostalih. Predstavi podatkovno rudarjenje in razloži primer uporabe podatkovnega rudarjenja z analizo podatkov kartic zvestobe (z njimi se lahko veliko uporabnikov poistoveti in problemsko razumejo domeno). Orange nima namenskega promocijskega videa. V naslednjem poglavju bo opisano ustvarjanje dveh videov, ki predstavijo podatkovno rudarjenje in orodje Orange začetnim uporabnikom.

Parametri/orodje	RapidMiner	KNIME	Weka	Orange
Št. video posnetkov na YouTube kanalu	109	172	91	9
Št. predvajalnih list na YouTube kanalu	16	29	4	1
Vodenje skozi izobraževalne vsebine po prenosu programa	Da	Da	Ne	Ne
Videi specifično narejeni, da si sledijo	Da	Ne	Da	Da
Skupno št. ogledov vseh posnetkov na YouTube kanalu	403,249	318,907	1,077,868	63,920

Tabela 2.1: Nekaj parametrov, povezanih z video vsebinami.

Oba bosta predstavila tako uporabo orodja kot tudi predstavitev problema in rešitve po zgledu promocijskega videa RapidMiner.

Orodje	Naslov	Dolžina	Opis
RapidMiner	RapidMiner Studio in 60 Seconds	1 min	Predstavi problem ugotavljanja verjetnosti preživetja na naboru podatkov udeležencev nesreče Titanika in njegovo rešitev.
KNIME	Welcome to KNIME TV	2min	Predstavi orodje preko razlage diapozitivov in slik uporabniškega vmesnika.
Weka	Data Mining with Weka: Trailer	5min	Predstavi, kaj je podatkovno rudarjenje in poda primer iz vsakdanjega življenja, kaj omogoča brez uporabe samega orodja.
Orange	/	/	/

Tabela 2.2: Kakšen je promocijski video vsakega od orodij?

Poglavje 3

Motivacijski videi za uporabo orodja Orange

Cilj naloge je bil ustvariti videe, ki bi širšemu občinstvu na preprost in komičen način predstavili proces pridobivanja znanja iz podatkov, njegove možnosti in izzive. Namen videov je ozaveščanje o tem področju, z uporabo elementov programa Orange pa želimo gledalce spodbuditi k uporabi tega orodja za začetek spoznavanja podatkovnega rudarjenja.

V sklopu pričujoče naloge sta nastala dva videa. Prvi predstavi rojenje (angl. clustering) [3] na naboru podatkov živali iz živalskega vrta[1], drugi pa pokaže gradnjo modela klasifikacijskega oz. odločitvenega drevesa [3] s podatki o treh vrstah cvetlic perunik (angl. Iris)[2].

3.1 Od scenarija do končnega videa

Za oba videa je bil najprej izbran nabor podatkov, ki smo ga želeli predstaviti v videu, glede na to pa je bila izbrana ključna lokacija snemanja (živalski vrt) ali pa predmet, okrog katerega bi potekal video (cvetlice perunike). Za vsak video je bilo sestavljenih več verzij scenarijev in popravkov, ki pa so v

[1] `https://archive.ics.uci.edu/ml/datasets/Zoo`
[2] `https://archive.ics.uci.edu/ml/datasets/Iris`

sodelovanju s sodelavci iz Laboratorija za bioinformatioko konvergirali v dva končna scenarija. Za vse scene smo določili končne lokacije, zbrali igralce in snemalno ekipo ter scene razdelali na kadre v snemalni knjigi. Temu so sledila usklajevanja snemalnih dni z igralci, snemalno ekipo in upravljavci živalskega vrta oz. dobavitelji cvetlic. Snemanji za oba videa skupaj sta bili opravljeni v skupnem trajanju treh snemalnih dni.

Posnetki so bili nato sestavljeni, dodana glasba in implementirani grafični elementi. Video je bil montiran v več iteracijah in izboljšan glede na pogovore z nekaj gledalci in mentorjem, zato se s scenarijem ne ujema v popolnosti. Nekatere scene so bile skrajšane ali pa celo v popolnosti odstranjene.

3.2 Video "Zoo"

Cilji videa je prikaz izbora najbolj informativnih atributov, ki so značilni za sesalce, ob tem pa predstaviti podatkovni nabor Zoo, prikazati proces zbiranja podatkov in kaj zmore podatkovno rudarjenje ob uporabi programa Orange.

3.2.1 Podatki

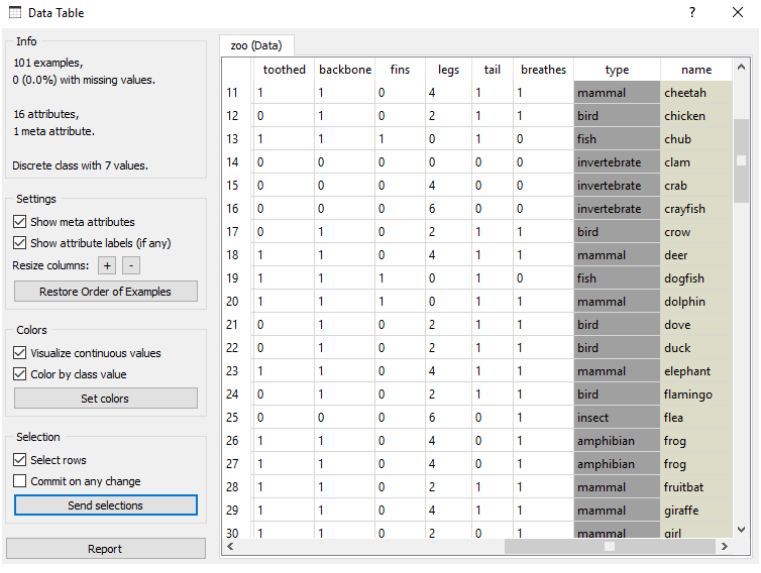
Podatkovni nabor, imenovan "Zoo", vsebuje 101 vnos brez manjkajočih vrednosti. Vsak vnos prestavlja eno žival z imenom in 16 drugimi atributi. Vrsto živali predstavlja razred, ki ima lahko eno izmed sedmih vrednosti. Te so vidne v tabeli 3.1, kot tudi porazdelitev zbranih živali v podatkovnem naboru glede na razred, na sliki 3.1 pa vidimo prikaz podatkovne tabele v orodju Orange.

3.2.2 Uporabljene metode strojnega učenja

Shema (slika 3.2) za obdelavo podatkov v videu vsebuje tri gradnike. Prvi, imenovan "File", zajame podatke iz datoteke, jih naloži v program in poda na razpolago za nadaljnjo analizo in obdelavo. Ta gradnik je povezano z

Razred	št. živali v razredu
sesalec (angl. mammal)	41
ptica (angl. bird)	20
plazilec (angl. reptile)	5
riba (angl. fish)	13
dvoživka (angl. amphibian)	4
insekt (angl. insect)	8
nevretenčar (angl. invertebrate)	10

Tabela 3.1: Porazdelitev živali v podatkovnem naboru "Zoo" glede na razred.



Slika 3.1: Izris nabora podatkov v orodju Orange preko gradnika za izris podatkovne tabele (angl. Data Table).

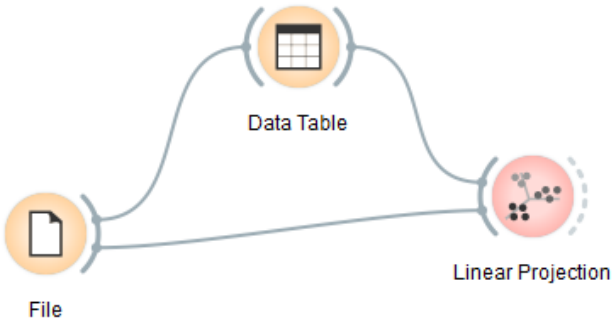
gradnikom “Data Table”, ki omogoča pregled in izbiro naloženih podatkov in vozliščem “Linear Projection”, ki podatke ne glede na njihovo dimenzionalnost preslika v dve dimenziji. Omenjena gradnika sta povezana tudi med seboj, kar v našem primeru omogoča pošiljanje izbranih podatkov iz gradnika “Data Table” v “Linear Projection” in jih tam prikaže tako, da so poslani elementi predstavljeni znotraj vizualizacije kot obarvani liki, medtem ko imajo ostali liki le obris.

V videu znotraj vozlišča za linearno projekcijo uporabimo algoritem “FreeViz”, ki je po besedah njegovih ustvarjalcev [4]:

> “... tehnika vizualizacije za analizo večdimenzionalnih podatkov, ki imajo določene vrednosti razredov. FreeViz vizualizacije lahko predstavijo podatke več parametrov v istem grafu, vendar ta preko optimizacijskega protokola izbere projekcijo, ki najbolje loči instance različnih razredov.”

To pomeni, da je izbrana prikazana projekcija glede na podane parametre najbolj optimalno ločila elemente različnih razredov oz. v našem primeru živali različnih vrst. Deluje kot metoda rojenja, saj poskuša oblikovati skupine elementov, ki so si znotraj skupine zelo podobni, izven skupine pa zelo različni [3].

Iz rezultata na sliki 3.5 vidimo, da so se glede na vrednosti atributov po razredih oblikovale skupine, ki so v nekaterih primerih, kot so ribe ali pa sesalci, bolj strnjene, v drugih primerih, kot so nevretenčarji ali pa plazilci, pa nekoliko bolj razpršene. Bolj strnjene skupine elementov so pozicionirane na skrajnih robovih vizualizacije ob oseh, ki predstavljajo za to skupino najbolj značilen atribut. Bolj razpršeni se nahajajo bližje izhodišču projekcije, kar bi lahko pomenilo, da med atributi, ki smo jih dodelili živalim, ne obstajajo tisti, ki bi bili izrazito značilni za to specifično skupino. V primeru sesalcev so se za parametre, ki to skupino najbolje opisujejo, izkazali zobje, mleko in dlaka.



Slika 3.2: Shema vozlišč v orodju Orange, uporabljena v videu "Zoo".

3.2.3 Predstavitev videa

V živalskem vrtu ob bazenu, vidnem na sliki 3.3, dva prijatelja opazujeta morskega leva, ko se eden vpraša, zakaj morski lev spada med sesalce. Odloči se, da vzrok poišče sam, zato se odpravi zbirat podatke o živalih v živalskem vrtu.

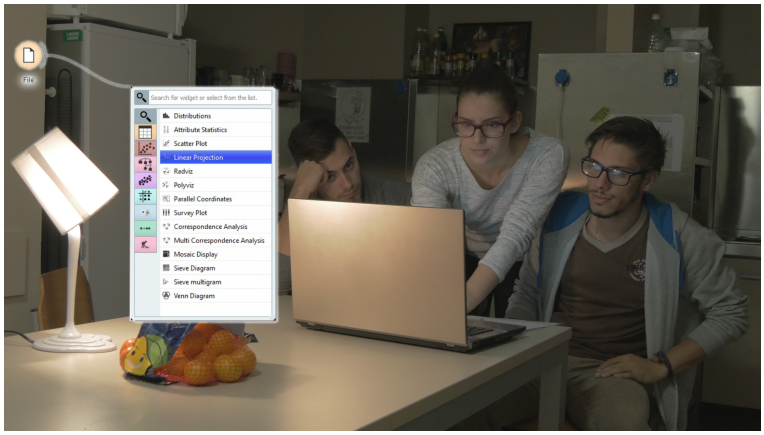
Z zbranimi podatki pristaneta doma v kuhinji in poskušata ugotoviti, kako bi iz vseh teh podatkov dobila odgovor na zastavljeno vprašanje. Ob razmišljanju se k njima prikrade cimra, ki podatke vnese v program Orange. Kot vidimo na sliki 3.4, odpre gradnik, ki zbrane podatke vizualizira v dvodimenzionalni graf in zažene optimizacijo.

Točke, ki predstavljajo živali, se zberejo v gručah v smeri parametrov, ki jih najbolje določajo, prijatelja pa lahko iz vizualizacije na sliki 3.5 razbereta, kateri parametri najbolje opredelijo skupino sesalcev.

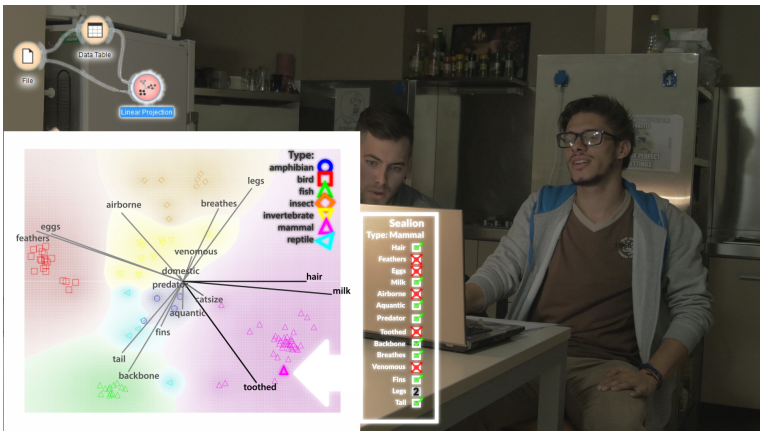
3.3 Video "Iris"

Cilj videa je prikaz gradnje razumljivega klasifikacijskega modela, ki ga lahko človek uporabi za klasifikacijo cvetlic perunik, ob tem pa tudi predstavitev podatkovnega nabora "Iris", prikaz procesa pridobivanja znanja ter prikaz

Slika 3.3: Glavna lika opazujeta morskega leva v njegovem bazenu.



Slika 3.4: Cimra sestavlja diagram poteka (angl. workflow) v programu Orange.



Slika 3.5: Glavna lika opazujeta podatke v projekciji in ugotavljata, da so parametri, ki najbolje predstavljajo sesalce, mleko, dlaka in zoblje.

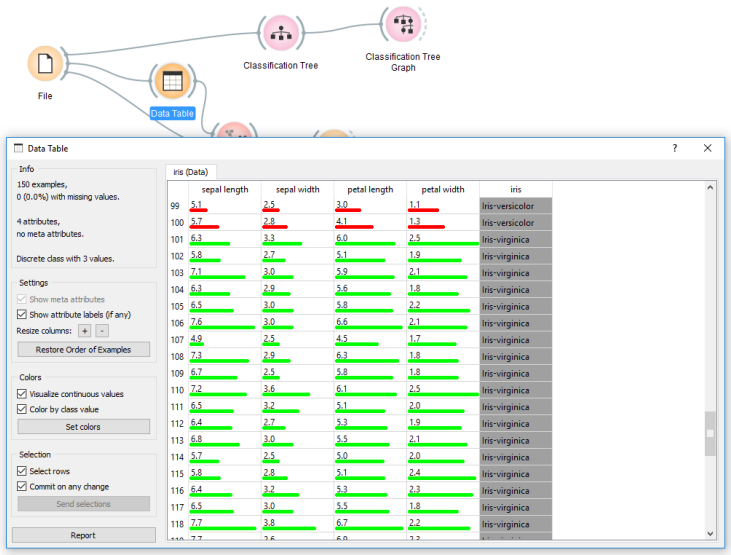
uporabe orodja Orange pri koraku podatkovnega rudarjenja.

3.3.1 Podatki

Podatkovni nabor imenovan "Iris" vsebuje 150 vnosov cvetlic brez manjkajočih vrednosti, vsak vnos pa je predstavljen s štirimi atributi:

- dolžina čašnega lista (sepal length)
- širina čašnega lista (sepal width)
- dolžina venčnega lista (petall length)
- širina venčnega lista (petal width)

Vsaka cvetlica ima določeno vrsto. Ta predstavlja razred, ki ima lahko eno izmed treh vrednosti (Setosa, Versicolor, Virginica). Cvetlice v podatkovnem naboru so enakomerno razdeljene med razredi, kar pomeni, da vsak razred vsebuje 50 vnosov. Na sliki 3.6 vidimo prikaz tabele v orodju Orange.



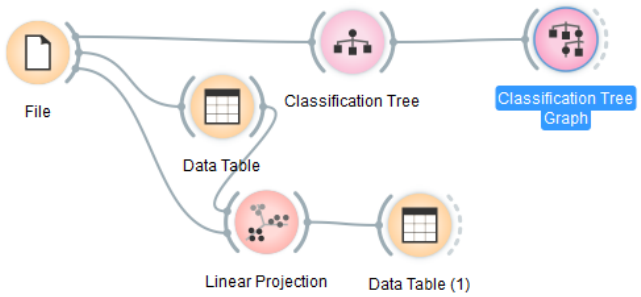
Slika 3.6: Izris nabora podatkov v orodju Orange preko komponente imenovane podatkovna tabela (angl. Data table).

3.3.2 Uporabljene metode strojnega učenja

V videu uporabljena shema za obdelavo podatkov, vsebuje dva dodatna gradnika glede na video "Zoo" (z izjemo gradnika "Data Table (1)", ki le izpiše izbrane podatke v linearni projekciji). "Classification Tree", ki omogoča gradnjo modela klasifikacijskega drevesa in nastavitev njegovih parametrov, izhod tega gradnika pa je povezan na "Classification Tree Graph", ki izriše zgrajeno drevo v obliki grafa, omogoča pa tudi nekaj vizualnih nastavitev.

V videu znotraj gradnika za gradnjo modela klasifikacijskega drevesa uporabimo informacijski prispevek (angl. information gain) kot kriterij za izbor atributov, izklopimo binarizacijo, v sekciji pred obrezovanjem drevja (angl. pre-pruning) omejimo deljenje vozlišč, ki vsebujejo večino razreda na 95 % in v sekciji obrezovanja drevesa po izgradnji (angl. post-pruning) omogočimo združevanje listov z istim večinskim razredom ter obrezovanje drevesa z m-oceno, kjer je "parameter m" enak 2.

Sama gradnja klasifikacijskega drevesa s pomočjo kriterija informacijskega



Slika 3.7: Shema vozlišč v orodju Orange uporabljena v videu "Iris".

prispevka poteka tako, da v vsakem vozlišču preverimo, ali vsi elementi pripadajo istemu razredu (ali vsaj 95 % razreda - pred obrezovanje drevesa). V nasprotnem primeru moramo poiskati atribut, ki v danem primeru najbolje loči vrednosti v čim bolj čiste skupine (vsebujejo le ali pa večinoma elemente istega razreda). Ker imamo v našem primeru atribute z zveznimi vrednostmi, je potreben še dodaten korak iskanja vrednosti, ki najbolje loči elemente na dve skupini. Ko je drevo zgrajeno, nastopi še obrezovanje drevesa, kjer se združijo vozlišča znotraj neke veje v list, ki predstavlja najpogostejši razred [3].

Iz rezultata na sliki 3.10 vidimo, da se je glede na vhodne vrednosti atributov oblikovalo drevo s 6 listi in 5 odločitvenimi vozlišči. V primeru uporabe tega odločitvenega drevesa smo ugotovili, da širina in višina čašnega lista nimata vpliva pri razlikovanju med cvetlicami in da cvetlice s širino venčnega lista manjšo ali enako 0,8 cm spadajo pod vrsto Iris Setosa z verjetnostjo 100 %. Klasifikacija ostalih dveh vrst je nekoliko bolj kompleksna, zato je predstavljena v tabeli 3.2.

Razred	širina venčnega lista	dolžina venčnega lista	verjetnost
Setosa	manjša ali enaka 0,8 cm	/	100%
Virginica	večja od 1,75 cm	/	97,8%
Virginica	med vključno 1,55 cm in vključno 1,75 cm	večja od 4,95 cm	100 %
Virginica	med vključno 1,55 cm in vključno 1,75 cm	večja od 5,45 cm	100 %
Versicolor	med 0,8 cm in vključno 1,75 cm	manjša ali enaka 4,95 cm	97,9 %
Versicolor	med vključno 1,55 cm in vključno 1,75 cm	med 4,95 cm in vključno 5,45 cm	100 %

Tabela 3.2: Povzetek pogojev in verjetnostna porazdelitev razreda v posameznem listu, glede na prikaz zgrajenega klasifikacijskega drevesa.

3.3.3 Predstavitev videa

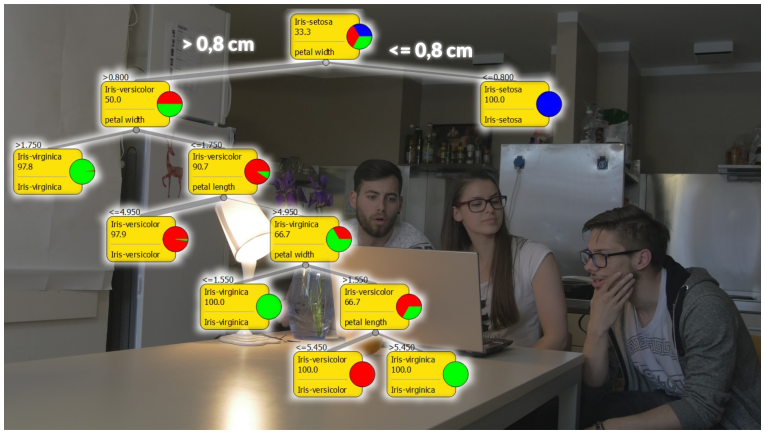
Ob prihodu domov eden od naših glavnih likov vidi drugega med opazovanjem cvetlic. Preko njunega pogovora na sliki 3.8 ugotovimo, da je ta kupil tri perunike, ki naj bi bile različnih vrst, a je pozabil, katera je katere vrste. Sedaj poskuša najti razlike, a jih ne vidi.

Ob njima se pojavi vseveda cimra, ki se odloči, da bodo skupaj rešili zagato. Odpravijo se izvajat meritve na perunikah z oznakami vrste (to je vidno na sliki 3.9).

Ko zberejo 150 vnosov, se s podatki vrnejo v kuhinjo in jih poskušajo analizirati enako, kot so to storili prejšnjič s podatki o živalih. Ker to ne odgovori na njihovo vprašanje, cimra doda par novih gradnikov in ustvari model odločitvenega drevesa. Z vizualizacijo odločitvenega drevesa, ki je vidna na sliki 3.10, so lahko določili vrsto vseh treh perunik, ki naj bi bile različne, in ugotovili, da so v resnici iste vrste, imenovane Setosa.

Slika 3.8: Desni lik sprašuje levega, zakaj opazuje perunike.

Slika 3.9: Izvajanje meritev na perunikah z oznakami vrste.



Slika 3.10: Odločitveno drevo, zgrajeno na podlagi zbranih podatkov o perunikah.

Poglavje 4

Sklepne ugotovitve

V nalogi smo analizirali spletne izobraževalne vsebine štirih najbolj priljubljenih odprtokodnih orodij za podatkovno rudarjenje z vmesnikom za vizualno programiranje. Poudarek je bil na video vsebinah, h katerim vsak od omenjenih pristopa na nekoliko drugačen način. Tako imajo pri RapidMinerju zelo sistematično razdelane vsebine za uporabo RapidMinerja, koncepte podatkovnega rudarjenja in primere, ki oboje združujejo. KNIME ima prav tako strukturirane vsebine, a so manj urejene tako na strani, kot tudi v obliki predvajalnih list, sestavljenih iz posnetkov, narejenih ob različnih priložnostih. Pri orodju Weka so se lotili izobraževanja zelo strukturirano, a za razliko od RapidMinerja z združenimi koncepti in uporabo orodja samega. Z uvedbo tečajev od učencev prejemajo informacije o dobrih in slabih točkah njihovega programa, kar jim omogoča izboljšave. Orange je na tem področju nov in se z vsebinami znajde nekje med RapidMinerjem in Weko.

Poleg analize sta nastala tudi dva predstavitvena videa za orodje Orange. Oba poskušata predstaviti podatkovno rudarjenje, njegove izzive in možnosti ter samo orodje Orange. To je uporabljeno tudi z namenom promocije uporabe programa za spoznavanje področja. V tej točki ni znano, ali je cilj teh dveh video predstavitev dosežen, saj v sklopu te naloge nismo izvedli strukturiranega testiranja z gledalci. To bi bil naslednji korak, potreben za izboljšanje teh dveh video predstavitev in pri razvoju naslednjih.

Iz analize orodij menimo, da je povezanost video vsebin zelo pomembna, saj so luknje in nejasnosti na naši strani v vlogi gledalcev pomenili nižanje pozornosti in volje do nadaljnjega učenja iz teh video vsebin.

Prav tako so se po našem mnenju odlično izkazale vodene vsebine v RapidMinerju, ki kombinirajo tekstovna navodila z grafičnimi znotraj samega orodja. Zanimivo bi bilo videti pristop, kjer bi bila interaktivnost še večja in bi vključevala vsaj avdio, če ne tudi video za vodenje skozi izobraževalne primere v samem programu. Dober primer tega so računalniške igre in mobilne aplikacije, tovrstna interakcija pa tudi odpira vrata za igrifikacijo (angl. gamification).

Literatura

[1] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[3] J. Han, J. Pei, and M. Kamber. *Data Mining, Southeast Asia Edition*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2006.

[4] Demšar J, Leban G, and Zupan B. Freeviz-an intelligent multivariate visualization approach to explorative analysis of biomedical data. *Journal of Biomedical Informatics*, 40(6):661–671, 2007.

[5] Alan Jovic, Karla Brkic, and Nikola Bogunovic. An overview of free software tools for general data mining. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1112–1117. IEEE, 2014.

[6] Robert A. Muenchen. The popularity of data analysis software. `http://r4stats.com/articles/popularity/`, 2016 (accessed October 16, 2016).

[7] Annan Naidu Paidi. Data mining: Future trends and applications. *International Journal of Modern Engineering Research (IJMER)*, 2(6):4657–4663, 2012.

[8] Gregory Piatetsky. Poll: What software you used for analytics, data mining, data science, machine learning projects in the past 12 months? `http://www.kdnuggets.com/2016/05/poll-analytics-data-mining-data-science-machine-learning-software-used.html`, 2016 (accessed October 15, 2016).

[9] Gregory Piatetsky. R, Python duel as top analytics, data science software – kdnuggets 2016 software poll results. `http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html`, 2016 (accessed October 15, 2016).

[10] Jason Wells, Robert Mathie Barry, and Aaron Spence. Using video tutorials as a carrot-and-stick approach to learning. *IEEE Transactions on Education*, 55(4):453–458, 2012.